CASE STUDY

Clinician AI Assist — Hospice System (clinicianaiassist.com)

A human-gated clinical AI suite for hospice, live in production — and the product-management record behind it, in full: Product Requirements · Evaluation Plan · Metrics Specification · Roadmap

The product	Six clinical-AI systems working one shared patient profile from referral to recertification — Hospice Analysis, Referral Audit, Plan Chart, Decline Trajectory, Clinical Audit, and Clinical Simulator — plus two profile-level risk instruments, Patient Chart Rating and Census Risk.
Status	Live in production at clinicianaiassist.com, with free sign-up. All suite models run gpt-5.6-luna. Two synthetic demo profiles — Dorothy A. Kessler and Robert J. Delgado — are public and read-only; every account-created profile is owner-scoped.
The builder	Kris Pueblos, RN, BSN — founder and product designer; Hospice Clinical Manager at a national provider (census ~40). He designed the systems and wrote their specifications, prototypes each system himself in Codex, and writes no production code: builders implement from his specs, and nothing ships without passing the verification gates he set.
Design rule	Human-gated by design: clinicians submit cases, independent validators and a final QA gate stand between any analysis and its reader, and the clinician decides. No prognosis, no eligibility or certification decisions, no autonomous orders.
Validation	A clinician testing round is recruiting now: practicing hospice physicians and RNs from the builder's own clinical network evaluate the suite on a synthetic testing kit, with structured evaluations collected on the product dashboard.
Date / evidence	November 9, 2026. Every figure in this case study is crosschecked against the governing systems spec (2026-11-05), the Patient Chart Rating model record (2026-11-06), and production records through 2026-11-08.

The case

Hospice care runs on documentation. Whether an agency is paid — and whether it withstands Medicare review — depends on whether the chart supports the care story: eligibility at admission, decline over time, the plan of care at recertification. The exposure rarely announces itself. A record can tell a good clinical story its documentation cannot support, and the agency's own standing with CMS is public data that sits disconnected from the chart in front of the clinician. When an Additional Documentation Request (ADR) arrives, the agency learns its exposure all at once.

Clinician AI Assist is a hospice decision-support suite built around a single judgment: **how much autonomy an AI system should get depends on the risk of its output**. Where the output touches a clinical record, the suite runs the strict model — nothing runs unattended, a clinician submits every case, independent validators and a final QA gate stand between any analysis and its reader, and the clinician makes every decision. The suite tracks trajectory; it predicts nothing. Two autonomous agents do exist in the product — an hourly operations rotation and a daily guideline watch — and they are fenced to data and structure only: they never run a clinical lane, never write a patient record, and never change a clinical output.

The systems exchange work only through governed handoffs: released outputs are origin-labeled and read-only to receiving systems, and the most recent data supersedes older data. Referral Audit and Plan Chart are start-of-care systems and take no inbound feeds — except Plan Chart recertification, which draws on all sources for its baseline. The other lanes draw on all systems.

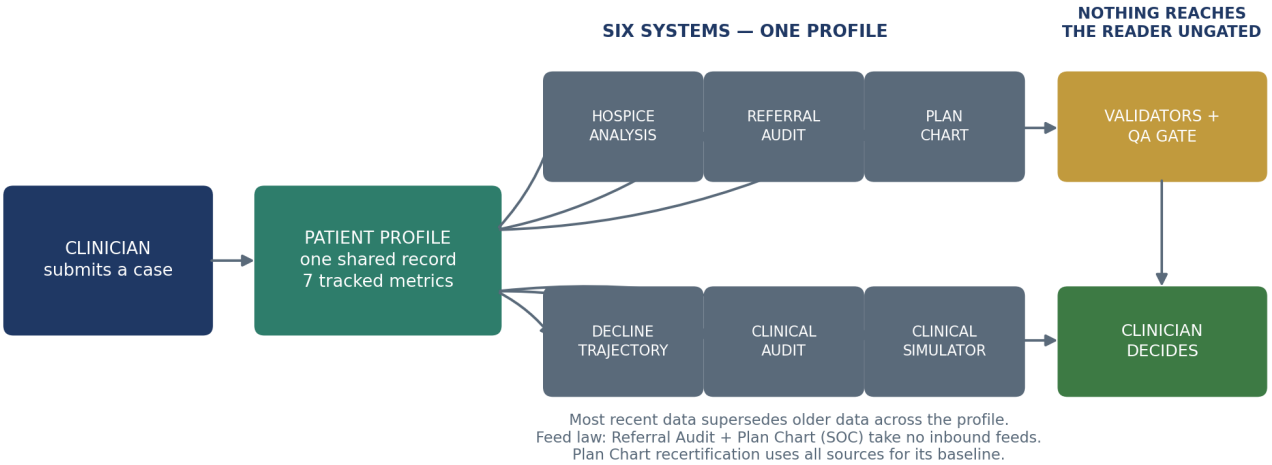


Figure 1 — Suite architecture and governance flow. Systems exchange work only through governed handoffs; released outputs are origin-labeled and read-only to receiving systems.

The six systems

System	What it does
Hospice Analysis	Hospice case analysis on the shared patient profile.
Referral Audit	Start-of-care system: audits a referral against its eligibility documentation and issues a disposition. That disposition feeds the appropriateness component of Patient Chart Rating (Part 1).
Plan Chart	Start-of-care system: reviews the plan of care and records a QA posture. Recertification reviews draw on all sources for the baseline.
Decline Trajectory	Reads documented decline over time (DAI/RDI). It tracks trajectory — it does not predict outcomes. Its reading feeds the justification component of Patient Chart Rating.
Clinical Audit	Per-visit deficiency scanning at maximum scrutiny — findings only, released under named dispositions. Finished and live; it is the subject of the evaluation plan in Part 2.
Clinical Simulator	Synthetic cases only — training and evaluation without touching a real record.

On the same profile sit **Patient Chart Rating** and **Census Risk** (Part 1): two cross-system instruments that join what the systems find in the chart to the agency's public CMS standing, so review exposure is visible early, in one place, with the full analysis shown behind every number.

The evidence so far (measured, from production records)

Measure	Value	Source
Systems live	6 of 6 (Clinical Audit finished; scrutiny mode only)	Production state, 2026-11-08
Clinical Audit — rebuilt reference audit	749 findings across 38 visits (High 158 · Moderate 470 · Low 118 · Watch 3); lane rating “Fair”; released with limitations	Stored audit output, 2026-11-07
Patient Chart Rating — demo patient	67 of 100 (Very High); Probable ADR likelihood 70 (Very High)	Live profile output, 2026-11-07
Census Risk — demonstration cohort	96 of 100 (Extremely High); probable multiple-ADR likelihood 64 (Very High)	Synthetic fixture run, 2026-11-07
Agency data coverage (SSVI)	6,642 of 7,059 hospices scored for FY2025; unscored reads “not published,” never zero	CMS SSVI file, Apr 2026 release
Stored Clinical Audit runs	12 of 12 scrutiny-mode and complete; zero correction-mode runs exist	Production records, 2026-11-08

Where it stands

The suite is feature-complete for its current purpose. Its open risk is not capability; it is external validation — and the roadmap (Part 4) is an evidence plan, not a feature list. The clinician testing round is the product's #1 priority: practicing hospice clinicians from the builder's own network work a synthetic testing kit — a case with three planted problems and a trajectory trap — and their structured evaluations land on the product dashboard. What follows is the working product record behind the suite, in full: the requirements document for its newest feature (Part 1), the evaluation plan for its highest-scrutiny system (Part 2), the metrics specification that governs the whole product (Part 3), and the roadmap those artifacts feed (Part 4).

Part 1 — Product Requirements Document: Patient Chart Rating

1.1 Problem

A hospice chart can look reassuring and still be exposed. The findings that matter under Medicare review — a stable vital sign, one good meal, one good transfer — are exactly the findings that mislead, and a record can tell a good clinical story its documentation cannot support. Separately, the agency's own standing with CMS is public data, but it sits disconnected from the chart in front of the clinician. When an Additional Documentation Request (ADR) arrives, the agency learns its exposure all at once. Patient Chart Rating exists to make that exposure visible early, in one place, with the full analysis shown behind the number.

1.2 Users and jobs

- Hospice clinical managers and IDG members: see which charts carry review exposure before CMS does, and why.
- Hospice medical directors: judge whether the record supports the care story at certification and recertification.
- Quality/compliance leads: connect chart-level findings to agency-level standing and aim QAPI work where it counts.

1.3 Goals and non-goals

- **Goal:** one combined rating per patient that joins chart evidence (four systems) with agency standing (public CMS data), every number shipped with its full analysis and a plain clinical translation.
- **Goal:** a Probable ADR likelihood per chart, and a census-level roll-up (Census Risk) with exposure priced by review level — MAC, SMRC, TPE, UPIC.
- **Non-goal:** no prognosis, no eligibility or certification decision, no order, no denial prediction. The product's language law: output states a **higher risk for elevation of investigation** — never fraud, never wrongdoing. PEPPER data is facility-confidential and is never modeled; it can only enter as an upload from the provider's own hands.

1.4 How the rating is computed

The model is deterministic — the same inputs always produce the same rating — and is computed at render from the most recent released output of each contributing system (the correlation law: most recent data supersedes).

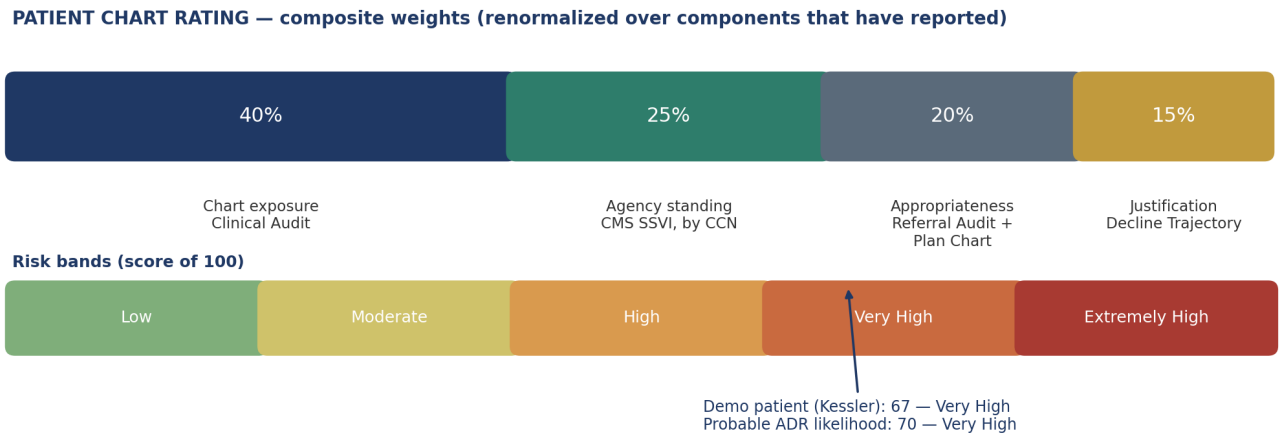


Figure 2 — Composite weights and risk bands, with the live demo patient's reading.

Component (weight)	Signal
Chart exposure (40%)	Latest Clinical Audit run. Deficiencies weighted by severity (Critical 10 · High 5 · Moderate 2 · Low 1 · Watch 0.5) per visit, on a saturating curve 100·d/(d+6); +8 when the technical share is at least 40%; floored at 70 when any fatal gap exists.
Agency standing (25%)	CMS SSVI by the profile's CCN: total/16 × 100, floored at 80 at 11+ and 92 at 13+ (CMS: 13+ ≈ top 1% of hospices). Layered with the public-data set keyed by CCN (PAC PUF, Hospice Care Index, HVLDL, HIS, CAHPS). Unpublished reads “not published,” never zero.
Appropriateness (20%)	Latest passing Referral Audit disposition (mapped 15–95; a hold floors at 75) averaged with the latest Plan Chart QA posture (20/40/75/90); then raised to at least 0.6 × chart exposure, because deficiency severity counts toward appropriateness.
Justification (15%)	Latest Decline Trajectory run: 100 – peak DAI, adjusted by RDI (FAILING –10 · FRAGILE +5 · STABLE +20 · IMPROVING +30), clamped 0–100. It runs backwards by design: a Justification of 0 means documented decline is maximally strong — the patient is actively failing — and the clinical translation says so explicitly.

Probable ADR likelihood combines selection pressure (SSVI 40% · outlier standing 30% · Hospice Care Index 15% · HVLDL 15%) with the chart's failure exposure as a geometric mean — an agency under scrutiny and a weak chart multiply each other; either one alone does not produce the top bands. **Census Risk** rolls

every rated chart into one census-level analysis with a probable multiple-ADR likelihood and severity priced by review level.

1.5 The translation requirement

Every number the rating shows also reads in plain clinical language, generated deterministically by the model itself — never written by hand after the fact. A “Guide — how to read the numbers” box opens the section (scale, band meanings, each component in plain words, the Justification-runs-backwards warning), and every headline number carries a generated “Why N” rationale at the number itself. The standing rule: translation **expands** the analysis; it never replaces or condenses it. A rating that rests on very few visits carries a concentration warning.

1.6 Acceptance criteria

- Deterministic: identical inputs reproduce the identical rating, bit for bit, on re-render and on the profile Re-analyze action (which realigns the correlated record and recomputes both risk sections in one click).
- Any component that has not reported is named on the profile as unmeasured evidence — the analysis is framed as “a better analysis as more data is introduced to the system,” never as partial or incomplete.
- Profiles are owner-scoped, verified at the database level; the only shared profiles are the two synthetic demo patients, read-only.
- Shareable output carries no real identifiers: the share infographic uses synthetic patient and agency names throughout, with no numbers, by design.

1.7 Status

Shipped and live on patient profiles. The demonstration patient (synthetic, end-stage ischemic cardiomyopathy) reads 67 of 100 (Very High) with a Probable ADR likelihood of 70 (Very High). Census Risk is live below it on the same profile pattern. Remaining work is evidence, not features: the clinician testing round (Part 4) is the current priority.

Part 2 — Evaluation Plan: Clinical Audit

Clinical Audit is the suite's highest-scrutiny system: per-visit deficiency scanning at maximum scrutiny, findings only, released under its own dispositions. Its scanning registry holds roughly 469 pattern-kind, 534 chain-kind, and 253 matrix-kind records. A system whose entire output is judgment — “this visit note does not support this claim” — cannot be evaluated with vibes. This plan applies Hamel Husain's method: **error analysis first, evals second** — you read real outputs, name the failure classes yourself, and only then build the measurements.

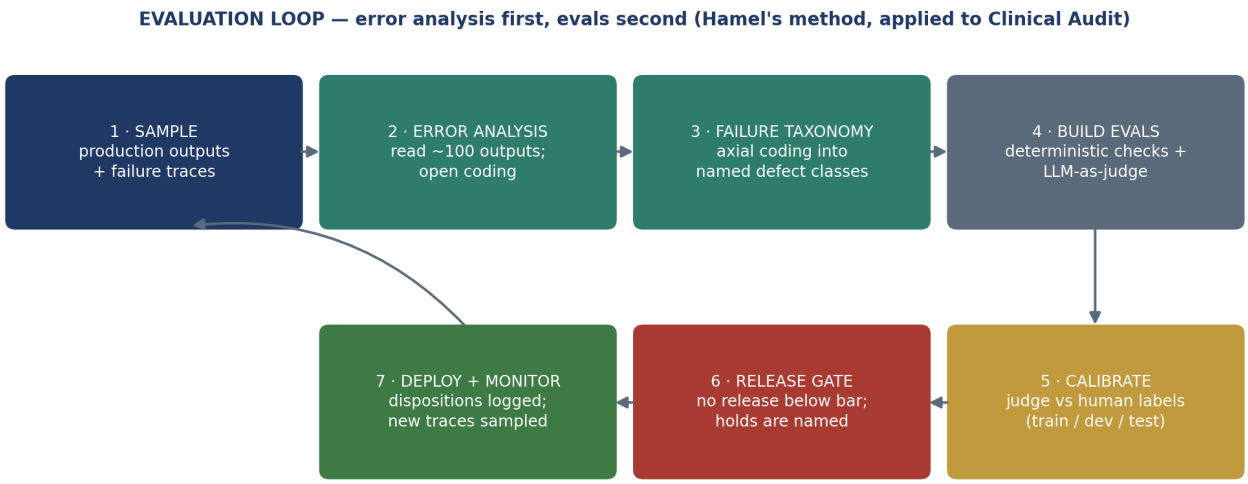


Figure 3 — The evaluation loop. The release gate is step 6: no run ships below the bar, and a hold is named, never hidden.

2.1 What error analysis has already found (production evidence)

- **Manufactured false stability.** Testing caught a defect in which a re-dated measurement made a declining patient read as stable. Root-cause fix: all five metric write paths were routed through a single validation gate. This class is now a zero-tolerance guardrail (Part 3).
- **Correction artifacts.** An earlier run form mixed audit with correction assistance. The mode was removed end-to-end (2026-11-08); the release gate now fails any run containing correction artifacts. All 12 stored runs are scrutiny-mode; zero correction runs ever existed.
- **Cross-account exposure.** A scripted cross-account audit before launch caught and closed real privacy policy holes. Profiles are owner-scoped at the database level; this check reruns on every schema or policy change.
- **Depth is load-bearing.** The rebuilt reference audit produced 749 findings across 38 visits. A full multi-visit chart audit takes about 25–30 minutes because the per-visit digest chain cannot be shortcut without losing findings; speed work is a roadmap item with a hard constraint, not a silent quality cut.

2.2 Building the eval suite (per failure class)

- **Binary criteria.** Every failure class gets a pass/fail criterion a reviewer can apply without interpretation — Hamel's rule: if two experts can't agree on the label, the criterion is rewritten, not debated.
- **Deterministic evals first.** Anything checkable in code is checked in code: severity weighting and band math, disposition logic, citation-to-source presence on every finding, correction-artifact absence, feed-law compliance (the audit receives documents only — never another system's clinical outputs as deficiency causes).
- **LLM-as-judge, calibrated.** For judgment calls (is this finding real? is the severity right?), a judge model scores outputs against criteria written from the error-analysis taxonomy. The judge is calibrated against expert labels — mine, as the domain expert annotator — on train/dev/test splits, and it may only gate releases once its agreement with the human labels on the held-out test split meets the bar in Part 3.
- **Golden sets.** The two synthetic demo patients carry fully processed records; their released outputs are frozen as regression fixtures. Any change to prompts, models, or the scanning registry must reproduce the expected findings profile on the fixtures before it can ship.

2.3 The release gate

Every run ends in a named disposition — **release**, **release with limitations**, or **hold** — issued after independent validators and the final QA step. Held work is shown as withheld, never silently dropped. The eval suite runs on any change to prompts, model version, or registry; a regression on a golden set or a guardrail class blocks the release, full stop.

2.4 Production monitoring and the flywheel

Dispositions, findings-per-visit distribution, and runtime are logged per run. New production outputs are sampled back into error analysis on a fixed cadence, so the failure taxonomy keeps growing from real traces instead of guesses. The clinician testing round (Part 4) feeds this loop with the highest-value traces of all: practicing hospice clinicians marking where the audit is wrong, harsh, or missing something.

Part 3 — Metrics Specification

North Star: Clinician-Validated Releases per Week — released analyses that a clinician opened, reviewed, and used (evaluation submitted, demo record inspected end-to-end, or pilot workflow completed). The suite's value is not analyses produced; it is analyses a clinician trusted enough to act on. Every other metric below either feeds that number or guards it. Targets are labeled TARGET; values already observed in production are labeled MEASURED.

Metric	Definition	Bar / status
Gate integrity	Share of released outputs that passed independent validators + QA with a named disposition; holds shown as withheld.	100% — MEASURED (structural)
Boundary violations	Outputs containing a prognosis, eligibility/certification decision, or order — the product's hard boundaries.	0 — guardrail, zero tolerance
False-stability class	Occurrences of the re-dated-measurement defect class across all metric write paths (single gate since the fix).	0 — guardrail, zero tolerance
Correction artifacts	Runs containing correction-assistance content (mode removed 2026-11-08; gate fails any such run).	0 of 12 stored runs — MEASURED
Judge–human agreement	LLM-as-judge agreement with expert labels on the held-out test split, per failure class, before the judge may gate.	TARGET: at least 90% per class
Audit depth	Findings per visit and severity mix per released audit, watched for silent thinning after any change.	Reference: 749 findings / 38 visits — MEASURED
Full-chart audit time	Wall-clock for a complete multi-visit chart audit (per-visit digest chain).	~25–30 min — MEASURED; reduce without breaking the chain
Rating coverage	Share of profiles with all four PCR components reported. Framing per the language law: the analysis gets better as more data is introduced.	TARGET: 100% on pilot census
Tester activation	Invited clinicians who complete at least one full evaluation on the testing kit.	TARGET: set at round launch (Part 4)
Clinical usability	Share of tester evaluations rating outputs clinically usable without edits that change meaning.	TARGET: at least 70% in round 1

Metric	Definition	Bar / status
Recommendations earned	LinkedIn recommendations from testers (1 physician, 1 RN) — part of the outreach gate.	TARGET: 2 by Jan 2027

Instrumentation note: disposition, coverage, and run-time metrics read from stored run records. Activation and usability metrics instrument with the testing round — the evaluation form is the collection point, and results land on the product dashboard at /evaluations.

Part 4 — Roadmap: evidence before applications

The suite is feature-complete for its current purpose. The roadmap is therefore not a feature list — it is an evidence plan. The product's open risk is not capability; it is external validation. Every phase below either produces that validation or converts it.

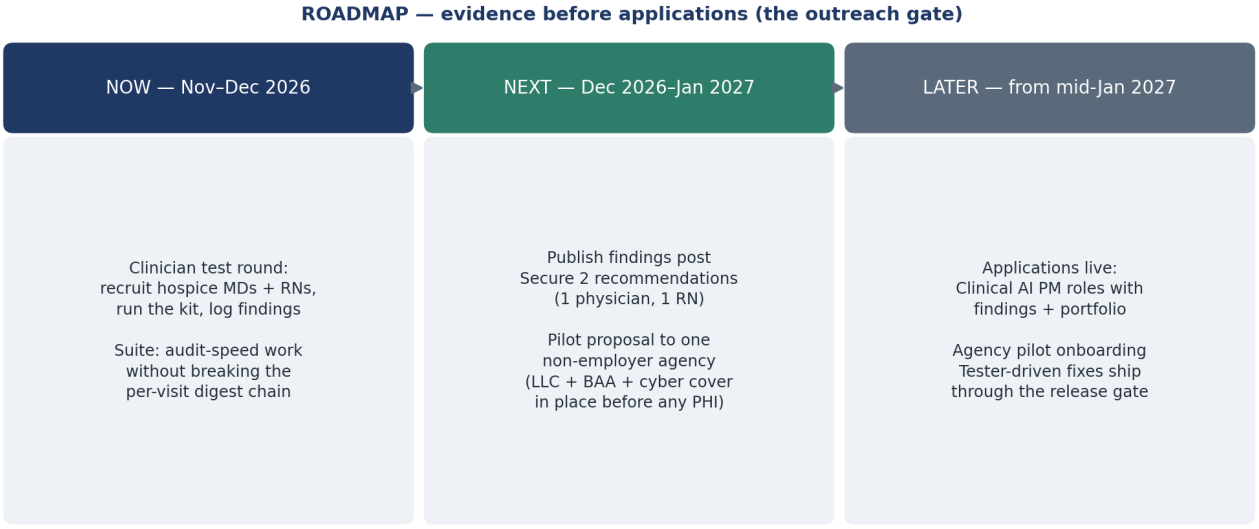


Figure 4 — Roadmap phases. The outreach gate: findings published + two recommendations + a quotable traffic snapshot from the companion venture + five rehearsed interview stories.

Now — Nov–Dec 2026

- Clinician testing round (the product's #1 priority): recruit hospice physicians and RNs from my own clinical network; run the testing kit (a synthetic case with three planted problems and a trajectory trap); collect structured evaluations on the dashboard.
- Audit-speed workstream: cut the 25–30 minute full-chart audit time with the digest chain intact; any speedup must hold findings-per-visit depth on the golden sets.

Next — Dec 2026–Jan 2027

- Publish the findings — including what the testers found wrong — as the suite's public evidence post; secure two recommendations (one physician, one RN).
- Pilot proposal (one non-employer hospice agency): scoped pilot with success criteria drawn from Part 3. Prerequisites before any real patient data: business entity, BAAs down the vendor chain, cyber insurance. My own employer is excluded by design.

Later — from mid-January 2027

- Applications live for Clinical AI Product Manager roles, leading with findings, recommendations, and this case study.
- Pilot onboarding; tester-driven fixes ship through the same release gate as everything else.

What is deliberately not on this roadmap: new systems, monetization claims, and any autonomous mode for clinical lanes. The governance model is the product's thesis; the roadmap protects it.

Appendix — sources and verification

Every figure in this case study was crosschecked against the system's own records: the governing spec CLINICIAN-AI-ASSIST-SYSTEMS-SPEC.md (2026-11-05) for system topology and feed law; PROFILE-RISK-MODEL-2026-11-06.md for the rating's inputs, weights, floors, and bands; production records through 2026-11-08 for run counts, the rebuilt audit's findings, the correction-mode removal, and model assignment (gpt-5.6-luna across the suite); and the CMS SSVI file (April 2026 release) for agency coverage counts. Demo patients are synthetic. Escalation wording follows the product's language law — “higher risk for elevation of investigation” — throughout. Companion venture (not part of this case study): Care and Living Guide, an autonomous multi-agent Medicare directory.